AgroLD: un graphe de connaissances pour la caractérisation des mécanismes moléculaires complexes impactant le phénome des plantes

Pierre Larmande[1,3], Gildas Tagny[2,3], Manuel Ruiz[2,3]

[1] UMR DIADE, IRD, Univ. of Montpellier, Montpellier, France.
pierre.larmande@ird.fr

[2] UMR AGAP, CIRAD, Montpellier, France
gildas.tagny_ngompe@cirad.fr, manuel.ruiz@cirad.fr

[3] SOUTH GREEN BIOINFORMATICS PLATFORM - Montpellier, France.

Résumé : La compréhension des interactions génotype-phénotype est un des axes les plus importants de la recherche en agronomie dont l'un des objectifs est d'accélérer la reproduction des caractères importants pour la production agricole. Or ces interactions sont complexes à identifier car elles s'expriment à différentes échelles moléculaires dans la plante et subissent de fortes influences de la part des facteurs environnementaux. Les technologies d'analyse haut-débit ne permettent de capturer que partiellement cette dynamique. Même si ces technologies sont de plus en plus performantes dans l'acquisition de données, notre connaissance du système reste encore parcellaire pour pouvoir comprendre les relations complexes existant entre les différents éléments moléculaires responsables de l'expression du phénome -ensemble des phénotypes observés pour un individu- . Cet objectif ne peut être atteint qu'en intégrant des informations de différents niveaux dans un modèle intégrateur utilisant une approche systémique afin de comprendre le fonctionnement réel d'un système biologique. Aujourd'hui, le Web sémantique propose des technologies pour l'intégration de données hétérogènes et leur transformation en connaissances explicites grâce aux ontologies.

Nous avons développé AgroLD (Venkatesan *et al.*, 2018) (Agronomic Linked Data - www.agrold.org), une base de connaissances reposant sur les technologies du Web sémantique et exploitant des ontologies du domaine biologique, afin d'intégrer des données issues de plusieurs espèces de plantes présentant un intérêt important pour la communauté scientifique, comme par exemple le riz, le blé et arabidopsis. Nous présentons les résultats du projet, qui portait initialement sur la génomique, la protéomique et la phénomique. AgroLD est aujourd'hui une base de plus de 100 millions de triplets créée à partir de plus de 50 jeux de données provenant d'une dizaine de sources de données, telles que Gramene (Tello-Ruiz *et al.*, 2018) et TropGeneDB (Hamelin *et al.*, 2012). Par ailleurs, nous avons utilisé une dizaine d'ontologies du domaine biologique, telles que Gene Ontology (The Gene Ontology Consortium, 2014) et Plant Ontology (Plant & Consortium, 2002) pour annoter et intégrer ces ressources. Pour cette phase, chaque jeu de données a été transformé à partir de sources sélectionnées et annotées sémantiquement en réutilisant les champs textuels correspondant avec des termes d'ontologies lorsqu'ils ont été fournis par la source d'origine. De plus, nous avons utilisé les services Web d'AgroPortal (Jonquet *et al.*, 2018) pour annoter sémantiquement des éléments supplémentaires tels que par exemple, les URIs correspondant à la taxonomie des espèces ou des éléments d'anatomie. Dans ces cas, nous avons généré des propriétés supplémentaires à partir des ontologies correspondantes, ajoutant ainsi 22% de triplets supplémentaires qui ont été validés manuellement.

L'objectif d'AgroLD est d'offrir une plate-forme de connaissances spécifiques du domaine agronomique afin de répondre à des questions biologiques complexes. De telles questions peuvent concerner le rôle de gènes spécifiques dans les mécanismes de résistance aux maladies des plantes ou de caractères de production identifiés à partir des analyses GWAS. Afin de rendre AgroLD accessible par un plus grand nombre d'utilisateurs, nous avons également développé une application Web proposant plusieurs interfaces de requêtes. Tout d'abord une interface simple qui permet aux utilisateurs de rechercher par mots-clés sur l'ensemble des valeurs de la base et ainsi de parcourir le contenu d'AgroLD. Puis une interface de recherche avancée qui permet de combiner du texte libre et des filtres à facettes ainsi que des services Web externes proposant ainsi une interface d'agrégation de données distribuées. AgroLD possède également une interface de visualisation des graphes qu'il est possible de configurer pour mettre en valeur certains types de relations. Finalement, un éditeur SPARQL propose un environnement interactif pour formuler des requêtes et manipuler des résultats. Actuellement, de nouveaux jeux de données sont en cours d'intégration. Ils portent sur les réseaux d'interaction protéine-protéine, les facteurs de transcription et réseaux de co-expression afin d'étendre les connaissances sur les mécanismes moléculaires. De nombreux développements sont également réalisés au niveau des interfaces de requêtes, notamment au niveau de la visualisation des graphes afin de fournir des outils plus dynamiques, interactifs et contextualisés. Enfin, une attention particulière est portée sur la qualité des données intégrées. Des méthodes de liage et de machine learning sont développées pour rechercher des liens et des ressources similaires dans la base de connaissances ou dans des ressources externes.
Mots-clés : Base de connaissances, Web sémantique, Agronomie, Génomique fonctionnelle, Phénome

IC 2019

Références

HAMELIN C., SEMPERE G., JOUFFE V. & RUIZ M. (2012). TropGeneDB, the multi-tropical crop information system updated and extended. *Nucleic acids research*, p. gks1105.

JONQUET C., TOULET A., ARNAUD E., AUBIN S., DZALÉ YEUMO E., EMONET V., GRAYBEAL J., LAPORTE M.-A., MUSEN M. A., PESCE V. & LARMANDE P. (2018). AgroPortal : A vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture*, **144**, 126–143.

PLANT T. & CONSORTIUM O. (2002). The Plant Ontology Consortium and plant ontologies. *Comparative and functional genomics*, **3**(2), 137–42. Citation Key : Plant2002.

TELLO-RUIZ M. K., NAITHANI S., STEIN J. C., GUPTA P., CAMPBELL M., OLSON A., WEI S., PREECE J., GENIZA M. J., JIAO Y., LEE Y. K., WANG B., MULVANEY J., CHOUGULE K., ELSER J., AL-BADER N., KUMARI S., THOMASON J., KUMAR V., BOLSER D. M., NAAMATI G., TAPANARI E., FONSECA N., HUERTA L., IQBAL H., KEAYS M., MUNOZ-POMER FUENTES A., TANG A., FABREGAT A., D'EUSTACHIO P., WEISER J., STEIN L. D., PETRYSZAK R., PAPATHEODOROU I., KERSEY P. J., LOCKHART P., TAYLOR C., JAISWAL P. & WARE D. (2018). Gramene 2018 : Unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Research*.

THE GENE ONTOLOGY CONSORTIUM (2014). Gene Ontology Consortium : going forward. *Nucleic acids research*, **43**(D1), D1049–1056.

VENKATESAN A., TAGNY NGOMPE G., HASSOUNI N. E., CHENTLI I., GUIGNON V., JONQUET C., RUIZ M. & LARMANDE P. (2018). Agronomic Linked Data (AgroLD) : A knowledge-based system to enable integrative biology in agronomy. *PLOS ONE*, **13**(11), 1–17.